PDVN: A Patch-based Dual-view Network for Face Liveness Detection using Light Field Focal Stack

Yunlong Wang[1]*, Mupei Li[1]*, Zhengquan Luo[1,2], Zhenan Sun[1]
[1]Institute of Automation, Chinese Academy of Sciences (CASIA)
[2]University of Science and Technology of China (USTC)
yunlong.wang@cripac.ia.ac.cn

Abstract

Light Field Focal Stack (LFFS) can be efficiently rendered from a light field (LF) image captured by plenoptic cameras. Differences in the 3D surface and texture of biometric samples are internally reflected in the defocus blur and local patterns between the rendered slices of LFFS. This unique property makes LFFS quite appropriate to differentiate presentation attack instruments (PAIs) from bona fide samples. A patch-based dual-view network (PDVN) is proposed in this paper to leverage the merits of LFFS for face presentation attack detection (PAD). First, original LFFS data are divided into various local patches along spatial dimensions, which distracts the model from learning the useless facial semantics and greatly relieve the problem of insufficient samples. The strategy of dual-view branches is innovatively proposed, wherein the original view and microscopic view can simultaneously contribute to liveness detection. Separable 3D convolution on the focal dimension is verified to be more effective than vanilla 3D convolution for extracting discriminative features from LFFS data. The voting mechanism on predictions of patch LFFS samples further strengthens the robustness of the proposed framework. PDVN is compared with other face PAD methods on IST LLFFSD dataset and achieves perfect performance, i.e., ACER drops to 0.

1. Introduction

Biometric identification systems have been widely used and deployed in recent years. Face recognition has become one of the mainstream traits due to its advantages of non-invasive acquisition, high accuracy, good interoperability, and moderate equipment expense. Face liveness detection is an indispensable module of any face recognition systems that blocks spoof attacks from malicious entities. Despite the continuous development of face PAD methods, diverse PAIs are constantly evolving, which puts more pressure on guaranteeing the security of face recognition systems.

In the literature, software-based face liveness detection methods usually rely on subtle differences in texture [1, 2, 3], shape [4], color [5, 6], and context [7, 8] to distinguish bona fide and spoofed face samples. Recently, deep learning (DL) techniques working on 2D planar images have achieved promising results on certain datasets [9, 10, 11, 12, 13]. However, the generalization ability of these methods on unseen spoof attacks is quite limited. The defects lie in that the missing stereoscopic characteristics of facial regions in 2D projections make it intractable to consider the discrepancy in 3D geometric structure. On the other hand, extra equipment or add-on components could be utilized in hardware-based face PAD methods to analyze the inherent properties of the living faces. 3D depth sensors like structured light cameras can filter out 2D planar attacks with an accurate depth map [14]. Near-infrared (NIR) cameras can reveal obvious differences of PAIs made in light-emitting diode (LED) materials [15]. Thermal imaging also works in discovering temperature distribution on the human face [16]. Multi-modal databases for face anti-spoofing are also emerging, such as MLFP [17], WMCA [18], CASIA-SURF [19] and so on.

In contrast with structure light/NIR/thermal imaging devices, LF imaging is a passive depth sensing technology. With the aid of an internal micro lens array (MLA), LF cameras are able to capture 4D spatial-angular information in a single photographic exposure, which implicitly records 3D geometric structure and reflectance property of the object surface. Actually, the differences in 3D surface geometry and reflectance between bona fide and spoofed face samples are the intrinsic properties that face PAD methods can exploit. Flat face PAIs including printed papers, printed photos, and electronic displays exhibit limited distinctions in the depth layout of facial regions. PAIs that simulate 3D geometric structure are thus more challenging, such as wrapped papers, curved photos, and latex masks. However, these 3D-simulation PAIs inevitably transform the face tex-

*contribute equally

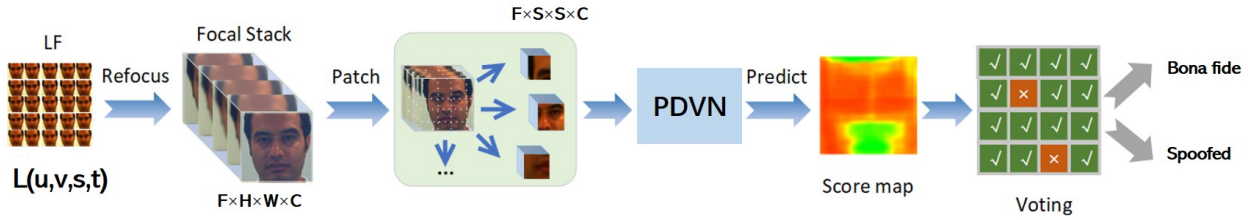

Figure 1. Overview of the proposed framework.

ture as a consequence of changes in the shadows and contrast gradients.

Several face PAD methods using LF imaging have been proposed recently. Most of these methods are based on specific handcrafted features analyzing the underlying cues in sub-aperture images (SAIs) or microlens images [20, 21, 22, 23, 24, 25]. More recently, convolutional neural network (CNN) was also adopted in LF-based face liveness detection [26], wherein the ray difference images and the microlens images are fed into CNNs to extract features.

LFFS can be efficiently rendered from a LF image captured by digital refocus. The defocus blur and local patterns between the rendered slices of LFFS reflect the differences in the 3D surface and 2D texture of biometric samples. In addition, LFFS data have the prominent convenience in compact data storage when compared with large raw LF images, dozens of SAIs, or hundreds of microlens images. Currently, there is nil work exploring the merits of both LFFS and CNN. In this paper, The proposed framework PDVN exploits the inherent features of LFFS data for face liveness detection instead of SAIs or microlens images.

2. Related Work

2.1. Traditional Face PAD Methods

Traditional image feature extraction methods were employed in the early stage such as LBP [1, 2], HOG [27], GLCM [3], DoG [7], SIFT [8], SURF [6], etc. These approaches mostly analyze subtle differences in texture patterns between bona fide face images and spoofed ones. The effect of color variations was also considered [5, 6], *e.g.*, converting face images to HSV and YCbCr color space to exploit implicit cues for binary classification. Other solutions based on image distortion [28] and Fourier spectra [29] were also investigated. In addition, face PAD techniques based on video clips could discover dynamic information like head/eye moving and expressions. For instance, DMD [30] aggregates video dynamics into one single image. Multi-LBP [31] adopts optical flow for motion estimation and encodes texture features into a multi-feature descriptor for video frames. Human facial physiology was also employed as liveness detection cues. Remote rhotoplethysmography (rPPG) was applied in face anti-spoofing relying on pulse estimation [32]. Facial expression and action also contribute to liveness detection such as eye blink [33, 34] and micro-motion analysis [35].

2.2. DL based Face PAD Methods

With the great success of DL in vision tasks such as object detection and recognition, CNN based methods were gradually introduced into face PAD solutions. Yang *et al*. [9] proved the effectiveness of CNNs in distinguishing face spoof attacks. Pre-trained CNN [10] could be regarded as deep feature extractors and combined with an SVM classifier to detect PAIs. Except for applying CNN to 2D images, the combination of CNN and long short term memory (LSTM) network structure was adopted in [36] to learn spatio-temporal features from videos for face anti-spoofing. Recently, some work modify the target of CNN training rather than the simple binary classification of bona fide and spoofed samples. In [11], face images are decomposed into two components, *i.e.*, real face and extra noise, equivalently converting face PAD task to noise modeling. The generalization ability of DL based methods on unseen PAIs is also a significant factor. [13] investigated the overfitting problems of face PAD models. [12] devoted to adjusting the features distribution of bona fide and spoofed samples. Novel network structures of PAD models also draw much attention. DTN [37] is one kind of tree-like network designed for unknown PAIs. In [38], an extreme lightweight network for face PAD is proposed to use only depth maps.

2.3. Face PAD Methods using LF Imaging

LF imaging gains huge advantages in exploring both surface and texture features for face PAD. GUC-LiFFAD [21] is the first publicly available database which shed light on using LF imaging in face PAD. IST LLFFSD [39] is widely acknowledged in this field, which contains 6 types of PAIs. Some handcrafted approaches have been proposed to adopt LF imaging for face PAD [20, 21, 22, 23, 24, 25]. Recently, CNN was also adopted in LF-based face liveness detection [26], using the ray difference images and the microlens images to extract features by CNNs. In similar tasks such as face recognition and emotion estimation, [40] leverages VGG as the backbone to extract deep features from the disparity and depth maps derived from LF image. SAIs decoded from raw LF data could be regarded as a frame sequence as the viewpoint translates. Hence, sequence mod-

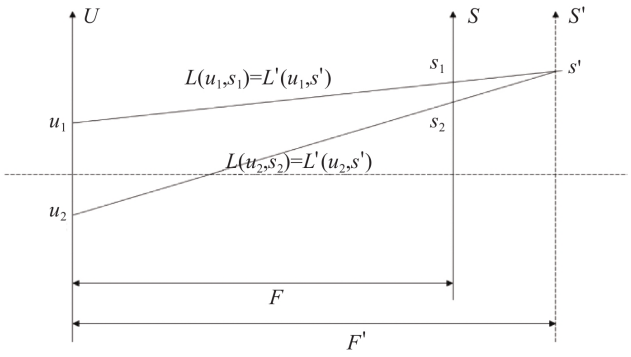

Figure 2. Schematic of digital refocusing. Only two dimensions (u, s) are displayed.

eling techniques were also utilized in face PAD using LF imaging. In [41, 42, 43], the models in the form of CNN-LSTM take the sequence of SAIs as input and outputs the feature aggregated from a set of viewpoints. These approaches demonstrate the potential of applying LF imaging to face PAD. However, scarce DL based face PAD frameworks take LFFS data as input.

3. Framework

An overview of the proposed framework is depicted in Fig. 1, and the details will be elaborated in this section.

3.1. Light Field Focal Stack

4D LF data can be decoded from LF images and expressed as the two-plane parameterization model $L(u, v, s, t)$. The traditional 2D planar image can be obtained by integrating the LF function as $I(s, t)$:

$$I(s,t) = \iint L(u,v,s,t) \cdot du \cdot dv \tag{1}$$

LF cameras can simultaneously capture the light intensity and directional information in a single exposure. Therefore, the focus plane can be changed at any depth layers by the digital refocusing method proposed by [44]. L' represents the synthetic film plane, and L is the original microlens plane. α indicates the relative locations of these two planes $\alpha = F'/F$. The schematic of digital refocusing is depicted in Fig.2, and can be derived as (2):

$$L'(u,v,s',t') = L\left(u, v, \frac{s'}{\alpha} + u(1-\alpha), \frac{t'}{\alpha} + v(1-\alpha)\right) \tag{2}$$

The focal stack reveals the 3D geometric structure of the imaged objects. Different facial region will exhibit different focusing levels as α varies. Even though the forged 2D texture details of spoofed faces can be similar to genuine

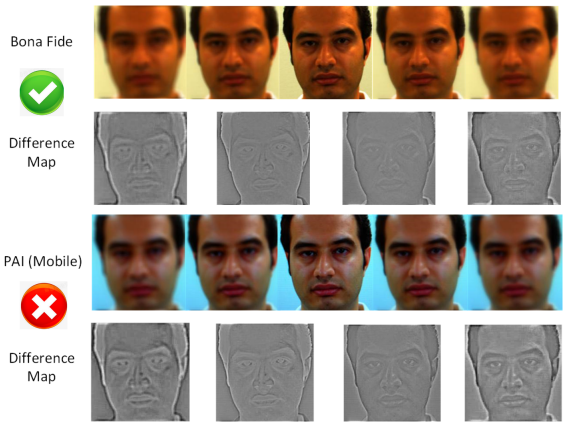

Figure 3. Exemplar slices of LFFS data. The first row is extracted from a bona fide LF sample, the second row is extracted from a spoofed LF sample presented on a mobile screen. The difference maps between the central slice and other neighboring slices are also shown, indicating the cues for liveness detection.

face samples, it is nearly impossible to fabricate the 3D geometric structure identical to real face physiology. On the other hand, the entities of current PAIs such as printed papers, printed photos, and electronic screens are usually flat or curved, which exhibit an almost constant amount of defocus blur between adjacent slices. The distinctions can be reflected in the difference between two adjacent slices in the focal stack as *Eq.* 3:

$$\begin{aligned} I_{\Delta\alpha}(s,t) &= I_{\alpha+\Delta\alpha}(s,t) - I_{\alpha}(s,t) \\ &\approx \iint -\Delta\alpha[(\frac{s}{\alpha^2}+u)\frac{\partial}{\partial s} + (\frac{t}{\alpha^2}+v)\frac{\partial}{\partial t}] \\ &\qquad L(u,v,\frac{s}{\alpha}+u(1-\alpha), \frac{t}{\alpha}+v(1-\alpha))]dudv \end{aligned} \tag{3}$$

Upon integral in the dimension of (u, v), it is theoretically verified in Eq.3 that the amount of pixel alteration differs as spatial location (s, t) varies. When the original LF image is in focus, the sharpest image is generated with $\alpha = 1$ in the focal stack. When α varies, the plane of refocusing image will deviate from the optimal imaging position, making the image blurry, which is shown in Fig.3. According to Eq.3, the defocus blur inside the focal stack largely depends on $\Delta\alpha$, denoting the step offset of the focus positions. Using different $\Delta\alpha$ will generate LFFS with different focal length in the focal dimension, which will significantly affect subsequent face PAD task.

3.2. Patch LFFS Data Generation

Unlike analyzing the geometrical layout of facial parts in recognition tasks, face PAD model strives to grasp the local differences in texture details and defocus distribution. In this paper, original LFFS data is proposed to be divided into

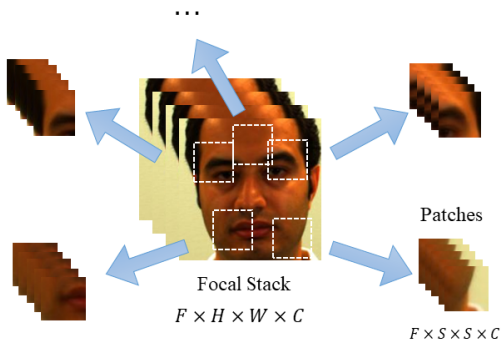

Figure 4. Patch LFFS data generation.

various local patches. LFFS obtained by digital refocusing is of the size $F \times H \times W \times C$. F is the length of the focal stack sequence, H and W are the height and width of each refocused slice. C is the number of channels which is generally 3 owing to the RGB color format. Next, random cropping is performed in a 2D spatial region. A square region cropped from a specific spatial location in each refocused slice is preserved, which will get patch LFFS data of the size $F \times S \times S \times C$. As shown in Fig.4, the spatial dimensions of patch LFFS data are cropped from a small portion of the whole face. The size of square region S generally depends on the requirements of training data and model complexity. To enrich the training data, different spatial locations can be randomly selected and the process of generating patch LFFS data can be repeated multiple times. Note that the labels of these generated patch data are all consistent with the original LF image.

Compared with complete LFFS data, the strategy of local patch input at arbitrary positions drives the PAD model to concentrate more on the intrinsic cues for liveness detection, rather than the layout of facial structure. Moreover, the number of LF images in the public domain is far from sufficient to train data-hungry DL frameworks for face liveness detection. Generating patch LFFS data exponentially increases the number of samples for training, which relieves insufficient samples of the whole face scale.

3.3. The Proposed Network

After patch LFFS data generation, the training samples and their corresponding labels are obtained. The proposed network is described in this section, which is used to predict whether one patch LFFS sample comes from bona fide LF data or not. The overall architecture of the proposed network is depicted in Fig.5.

3.3.1 Dual-view Branches

Semantic contexts are usually not the key to PAD solutions as photorealistic faces displayed by PAIs look quite similar to real ones. Instead, the details in the local region may reveal the devil of PAIs, including color variation, texture, edge and so on.

Taking a look at Fig.3 again, local patterns along spatial and focal dimensions of LFFS data exhibit different characteristics. Even in a small patch, the defocus level may also show a different tendency. This observation inspires us to propose a dual-view branches structure, wherein the original view and microscopic view can simultaneously contribute to finding spoofing cues for liveness detection. Specifically, the original view branch takes patch LFFS data as input and passes it to a sub-network containing four groups of separable 3D convolution. Separable 3D convolution on the focal dimension is elaborated in Sec.3.3.2. As to the microscopic view branch, the input is patch LFFS data with a much smaller spatial resolution, which is arbitrarily cropped from the input of the original view branch. This branch processes the micro-patch LFFS data with two groups of vanilla 3D convolution in [45]. Note that both branches adopt shortcut 3D convolution in each group of convolutional operations, which is functional as residual learning [46]. The intermediate feature maps of dual-view branches after convolutional layers are then aggregated into a 512d feature vector through global average pooling layers, respectively. These two feature vectors are then concatenated together to obtain a 1024d feature vector, which contains the complementary information of two branches. Its length is reduced to 64 via a fully connected (FC) layer and finally yields a single output node after another FC layer with a sigmoid activation function. The range of the output node is [0, 1], which means the confidence level that the patch sample is bona fide or not. The closer to 1, the more likely it is to predict that this patch comes from a bona fide face. Conversely, approaching 0 indicates that the patch is from PAIs. The details of the network parameters are marked in Fig.5.

3.3.2 Separable 3D Convolution on Focal Dimension

As stated above, patch LFFS data is 3D multi-channel tensor data with fixed spatial resolution, which therefore could be processed with 3D convolution [45]. 3D convolution is similar to traditional 2D convolution, operating by sliding the convolutional kernel smoothly in all directions, but the difference is that the input, output and kernel are all 3D tensors. Vanilla 3D convolution is suitable for sequence modeling, *e.g.*, video-based action recognition. However, the focal dimension of LFFS data is quite different from the temporal dimension of video clips. To be concrete, the field of view (FoV) of slices in LFFS is strictly aligned, but defocus patterns vary from each other in local areas due to various surface depths of facial parts. If vanilla 3D convolution is directly used for feature extraction, it may introduce ambiguities when convolving along the focal dimension. There-

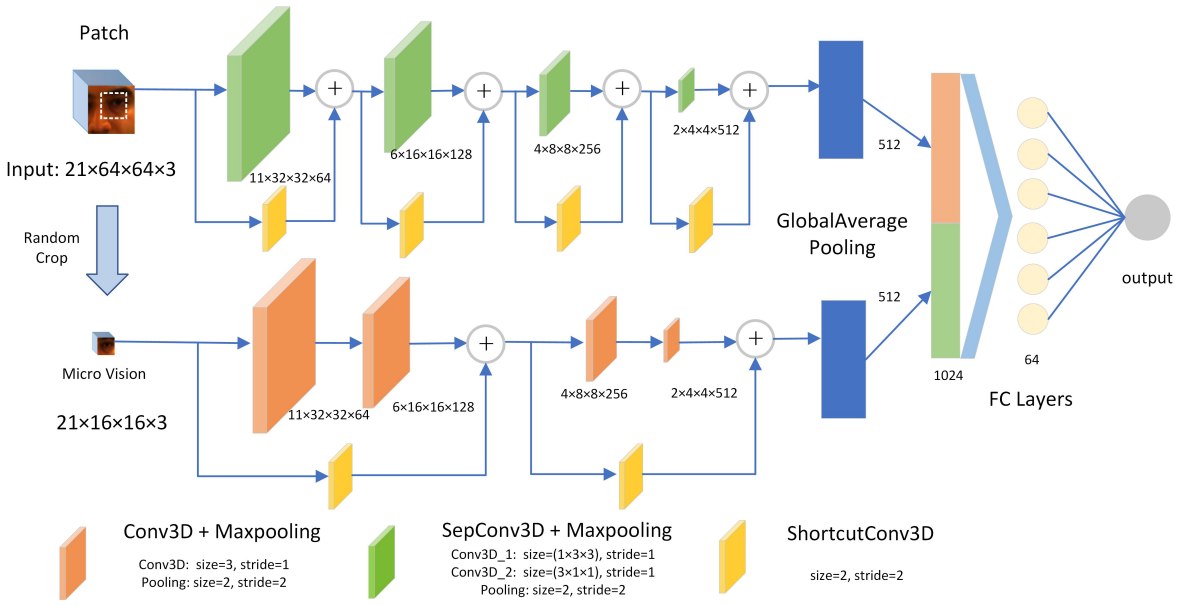

Figure 5. The overall architecture of the proposed network.

fore, an appropriate separable 3D convolution on the focal dimension is adopted as shown in Fig.6.

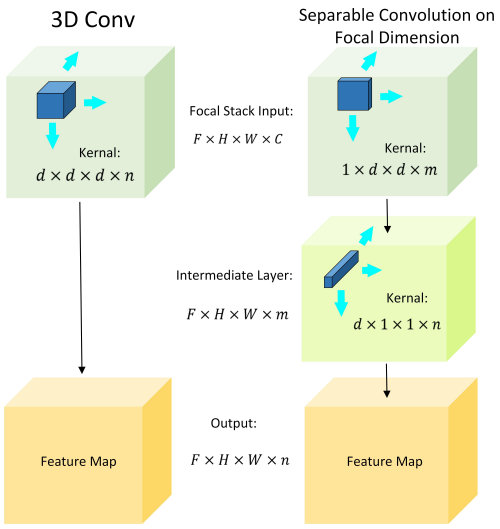

Figure 6. (a) Vanilla 3D convolution. (b) Separable 3D convolution on focal dimension.

In processing patch LFFS data, vanilla 3D convolution with a kernel size of $d \times d \times d$ is replaced by the following two-step consecutive convolution operations in the spatial dimensions and focal dimension. The first step is to use a 3D convolution with kernel size $1 \times d \times d$. Since the kernel size in the focal dimension is 1, only the spatial dimensions are convolved with $d \times d$. Second, a 3D convolution with kernel size $d \times 1 \times 1$ kernel is used to operate on the output feature map of the first step. Similarly, this operation is a one-dimensional convolution in the focal dimension. In this manner, the spatial texture features of each slice in LFFS are first extracted respectively, then intermediate representation at each pixel position among a set of adjacent depth channels are weighted and aggregated to explore the defocus variations. As demonstrated in Sec.4, experiments show that separable 3D convolution on the focal dimension has achieved better performance than vanilla 3D convolution methods.

3.4. Implementation Details

Refocus range S_{fs} and step offset $\Delta\alpha$. There is the following relationship.

$$S_{fs} = \Delta\alpha \cdot N_{fs} \tag{4}$$

where N_{fs} is the length of LFFS data. In the refocusing stage, the refocusing parameters are determined by an heuristic search method according to the blur-clear-blur phenomenon during refocusing. To be concrete, $\Delta\alpha = 0.0028$, $N_s = 21$.

Size of the cropped patch. In generating patch LFFS data for network training, the cropping size and number of patches need to be set manually. As for the original LF image in the IST LLFFSD [39] database, the resolution of the segmented face region is about 220×220 pixels. It is suitable to set the size of the cropped patch to about one-fifth of the face resolution, in which sufficient partial areas can be included and complete facial features will not be intercepted. Hence, the patch size of 64×64 and 16×16 microscopic vision area are appropriate, which also matches the network.

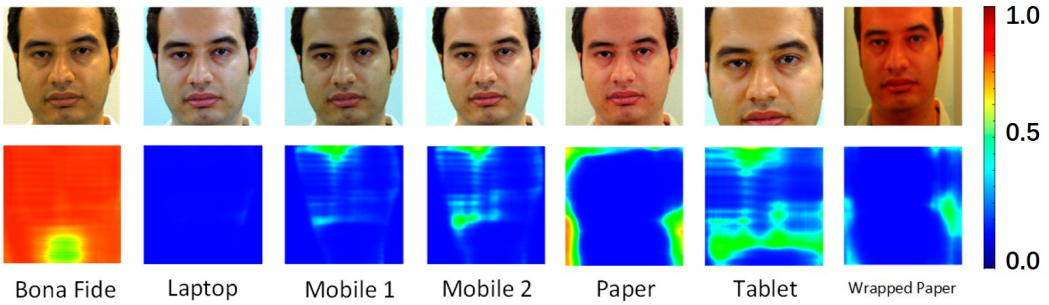

Figure 7. All predictions of patch samples constitute the score map of the original LF sample. Red (close to 1) means real and blue (close to 0) means spoofing.

Network training. The 16×16 microscopic vision area is arbitrarily selected inside the spatial patch during training. SGD optimizer is used for back-propagation. Warm-up training is adopted with an initial learning rate of 0.01 for 10 epochs. Thereafter, the learning rate is decayed by a factor of 10 every 5 epochs. The loss function is binary cross-entropy as Eq.5, where y is a ground-truth label (1 if a bona fide sample, otherwise 0) and $\hat{y}$ is the predicted probability. The model will converge after around 25 epochs using the MindSpore Lite tool [47].

$$loss = -y \cdot \log \hat{y} + (1 - y) \cdot \log(1 - \hat{y}) \tag{5}$$

Network inference. For original LF samples to be predicted, we first generate LFFS by digital refocusing. Then slide the cropping window at equal intervals to get patch samples, and randomly select the microscopic vision area in the center of each patch sample. After that, patch LFFS samples are fed into the proposed network to get predictions. All predictions of patch samples constitute score maps of the original LF sample as shown in Fig.7. Specifically, if the sliding step is 1 pixel, the resolution of the output score map equals to the resolution of original LF sample minus the spatial size of patch sample. Finally, the voting results of all patch samples are counted to determine the authenticity of the original LF sample for liveness detection. The number of patches should consider the accuracy required by the task and the complexity of the model. The more patches, the more stable the voting results will be and the less chance of misclassification, but the longer the computation time spent predicting the labels of these patches.

The final prediction result of the original LF sample is altogether determined by the prediction of each patch. When some local parts are obstructed by noises, a correct prediction result can still be obtained by considering the rest patches from clean parts. In other words, scattered local interferences can be significantly suppressed through this voting mechanism. In addition, the model's preference for predictions can be easily controlled by the voting ratio or confidence threshold.

4. Experiments

The dataset and metrics, experimental settings and results, and ablation study are presented in this section.

4.1. Dataset and Metrics

The proposed method and other compared methods were evaluated on the IST LLFFSD [39] database. The dataset consists of 50 subjects and 700 LF images. These LF images are categorized into a total of 100 groups, each of which contains 1 bona fide sample and 6 different PAI samples. These PAIs include laptop (*Lap*), tablet (*Tab*), mobile1 (*Mb1*), mobile2 (*Mb2*), paper (*Pap*) and wrapped paper (*Wpa*). 30 patch LFFS samples are generated from each LF image through random spatial cropping.

According to ISO/IEC 30107-3:2017 [48], the evaluation metrics include attack presentation classification error rate (APCER), bona fide presentation classification error rate (BPCER), and average classification error rate (ACER), which is the mean of APCER and BPCER. Detection Error Tradeoff (DET) curves are also plotted.

4.2. Experimental settings and results

Various liveness detection methods were compared on the IST LLFFSD dataset, including approaches operating on 2D images [1, 4, 28, 6], LF images [20, 21, 22, 39], and DL based methods. Note that the best-performing performances of these comparative methods are respectively adopted from the reported results in [39]. *2D CNN* is a ResNet50 [46] model which takes 2D central slice of LFFS data as input and outputs a 512d feature vector, *3D CNN* is a C3D model adopted from [45] which takes complete LFFS data as input and output a 512d feature vector. *2D CNN fusion* applies a ResNet50 model on each slice of LFFS data and concatenates all the extracted features, then reduces the feature dimension to 512d. *CNN-LSTM* first adopts a ResNet50 model to extract a 512d feature vector from each slice in LFFS data, then LSTM is used for sequence modeling and its output of last time step is regarded as the final representation. All these compared DL based methods

were retrained from scratch under the same dataset splitting as the proposed method. Without bells and whistles, no data augmentation tricks were used in training and the samples were processed in RGB color space. Bona fide and all PAI types were mixed and a 5-fold cross-validation was performed. The comparison results are shown in Table 1, and DET curves of different PAIs are shown in Fig.8.

	Lap	Tab	Mb1	Mb2	Pap	Wpa
2D image						
[1]	42.62	23.03	39.33	46.31	33.87	20.32
[4]	27.90	24.13	19.30	25.60	17.70	28.30
[28]	12.06	13.01	9.23	15.83	14.82	15.27
[6]	4.32	2.65	2.52	5.81	2.75	4.94
LF image						
[20]	10.12	12.39	12.79	13.86	12.91	16.14
[21]	19.78	26.36	29.98	22.46	32.43	38.03
[22]	11.00	10.77	8.12	18.50	7.27	22.05
[39]	0.88	2.14	0.73	0.79	0.75	2.85
DL methods (1) 2D CNN (2) 2D CNN fusion (3) CNN-LSTM (4) 3D CNN						
(1)	7.48	5.09	13.62	11.84	9.21	9.74
(2)	3.80	4.47	4.35	3.80	3.80	4.95
(3)	1.27	9.74	6.79	21.03	13.22	11.74
(4)	1.84	4.35	5.24	8.10	2.43	6.12
Ours	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Table 1. ACER of the proposed method and other compared PAD methods (%).

The proposed method achieves perfect performance and surpasses all the compared methods by a large margin. The ACER of all PAIs drops to zero, which demonstrates the superiority of the proposed framework. A major advantage of the patch-based scheme adopted in the proposed method is that the number of training samples increases exponentially by generating patch LFFS data. It is well known that only a sufficient amount of data samples can make the DNN model converge properly and fully learn the inherent features without overfitting. The IST LLFFSD dataset used in the experiment consists of only 700 LF samples but the number of patch samples can expand hundreds of times. Table 2 presents the model performances with different numbers of original LF images as training samples. It can be seen that the model trained with patch-based dataset expansion has a strong tolerance for dataset size. Even with only 4% samples (4 out of 100 groups) for training, ACER is as low as 7.81% when tested on the remaining 96% LF samples.

Training set size	4%	10%	15%	50%	80%
BPCER	6.25	5.55	1.25	0	0
APCER	9.37	3.33	0	0	0
ACER	7.81	4.33	0.62	0	0

Table 2. The model performances with different numbers of original LF images as training samples (%).

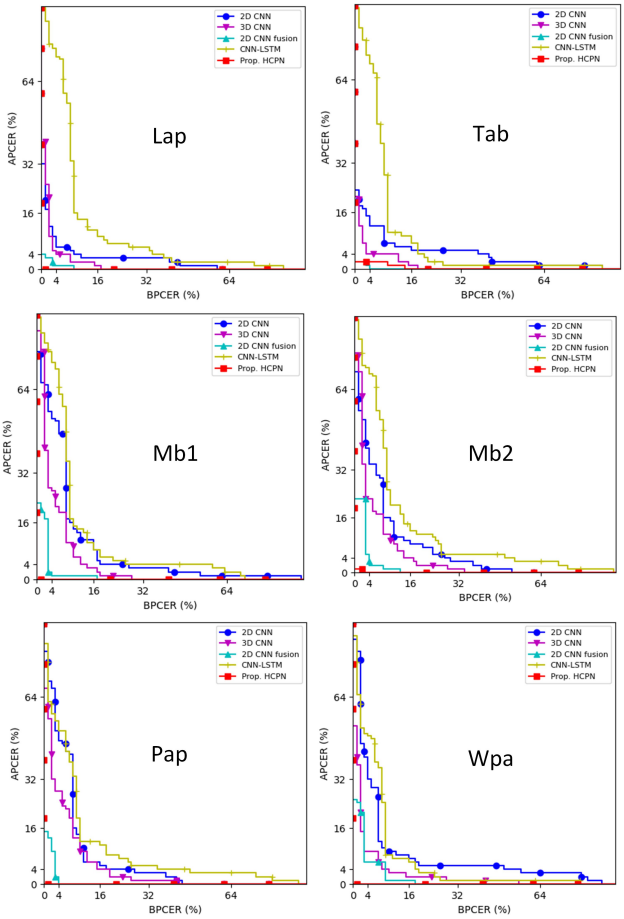

Figure 8. DET curves of DL based face PAD methods.

4.3. Ablation Study

4.3.1 Network Structure

To verify the effectiveness of the dual-view branches structure and separable 3D convolution on the focal dimension, ablation experiments are conducted. The complete framework here is denoted as *Dual-view*. Then, the micro-vision view is removed from *Dual-view* and a single branch with separable 3D convolution on the focal dimension is retained, denoted as *sep3D*. Next, we substitute separable 3D convolution in *sep3D* with vanilla 3D CNN [45] but maintain the same network structure, denoted as *vanilla3D*. The convergence speed, patch accuracy, and overall accuracy within 25 training epochs are depicted in Fig.9. It can be seen that the convergence speed of *Dual-view* is nearly the same as that of *sep3D*. Meanwhile, the patch accuracy and overall voting accuracy of *Dual-view* are slightly better than that of *sep3D*, and the reduction of error rate on the test dataset is more significant. Moreover, the convergence speed of *sep3D* is slightly slower than that of *vanilla3D*. It is because 3D convolutional operation in *sep3D* is decomposed into two consecutive layers. However, concerning the

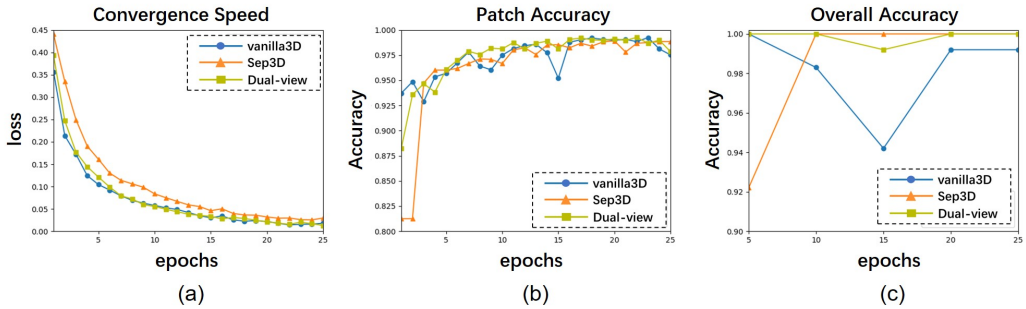

Figure 9. Ablation experiments of network structure. (a) Convergence speed. (b) Patch accuracy. (c) Overall accuracy.

same network structure, the amount of parameters of *sep3D* is less than that of *vanilla3D*. Moreover, the patch accuracy of *sep3D* is higher than that of *vanilla3D*, and the overall accuracy is also higher. Separable 3D convolution on the focal dimension is more powerful in extracting discriminative features from LFFS data than vanilla 3D convolution.

4.3.2 Influence of Refocusing Resolutions

The refocusing step offset $\Delta\alpha$ is determined heuristically. For different $\Delta\alpha$ in the same depth range, the resolution of LFFS data in the focal dimension will change. In terms of different refocusing resolutions, the performance fluctuation of the proposed method is investigated and compared with vanilla 3D CNN, and the results are shown in Table 3. It can be seen that the conditions for refocusing resolution are loose. When the refocusing step offset is enlarged by 2 or 3 times, ACER is on par with the original level.

	$\Delta\alpha$	$2\Delta\alpha$	$3\Delta\alpha$
3D CNN	6.81	4.74	6.77
Ours	0	0.12	0.68

Table 3. ACER of varying refocusing resolutions (%).

4.3.3 Impact of voting mechanism

In the proposed method, an important contribution to achieving the extremely low error rate lies in the voting mechanism of patch predictions. The final detection result is accumulated by the vote of each patch, and the conclusion is made on the side with more votes. Suppose that if the error rate of a single patch e is independent of the original data label and its neighboring regions, then the voting prediction error requires more than half of the patch predictions to be wrong, which makes the overall error rate E decrease exponentially. Fig. 10 shows how E varies with e and the number of patches. It can be seen that when using a low-accuracy patch model with a higher error rate, the overall voting error rate can be reduced by increasing the number of patches. However, if the patch error rate is too high, the effectiveness of using the voting mechanism is not significant.

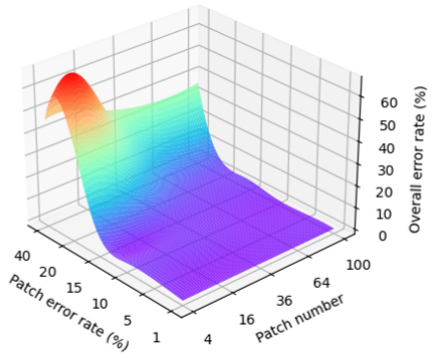

Figure 10. Curves of overall error rate E influenced by patch error rate e and patch number.

5. Conclusion

The proposed framework PDVN is validated to be a powerful DL based face PAD method using LFFS data, attributing to patch LFFS data generation, dual-view branches, separable 3D convolution on the focal dimension and voting mechanism on predictions of patch LFFS samples. Its superiority is experimentally verified on the widely adopted IST LLFFSD dataset. In future work, we will consider collecting databases and extending PDVN to other biometric traits, such as iris and fingerprint PAD.

6. Acknowledgement

This work is supported by the National Natural Science Foundation of China (Grant No. 62006225). We gratefully acknowledge the support of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References

[1] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single

images using micro-texture analysis. In *2011 international joint conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2011.

[2] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Lbp-top based countermeasure against face spoofing attacks. In *Asian Conference on Computer Vision*, pages 121–132. Springer, 2012.

[3] F Sthevanie and KN Ramadhani. Spoofing detection on facial images recognition using lbp and glcm combination. In *Journal of Physics: Conference Series*, volume 971, page 012014. IOP Publishing, 2018.

[4] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using texture and local shape analysis. *IET biometrics*, 1(1):3–10, 2012.

[5] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *2015 IEEE international conference on image processing (ICIP)*, pages 2636–2640. IEEE, 2015.

[6] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145, 2016.

[7] Bruno Peixoto, Carolina Michelassi, and Anderson Rocha. Face liveness detection under bad illumination conditions. In *2011 18th IEEE International Conference on Image Processing*, pages 3557–3560. IEEE, 2011.

[8] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *IEEE transactions on information forensics and security*, 11(10):2268–2283, 2016.

[9] Jianwei Yang, Zhen Lei, and Stan Z Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014.

[10] Lei Li, Xiaoyi Feng, Zinelabidine Boulkenafet, Zhaoqiang Xia, Mingming Li, and Abdenour Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2016.

[11] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *Proceedings of the European conference on computer vision (ECCV)*, pages 290–306, 2018.

[12] Xiaoguang Tu, Hengsheng Zhang, Mei Xie, Yao Luo, Yuefei Zhang, and Zheng Ma. Deep transfer across domains for face antispoofing. *Journal of Electronic Imaging*, 28(4):043001, 2019.

[13] Xiaoguang Tu, Zheng Ma, Jian Zhao, Guodong Du, Mei Xie, and Jiashi Feng. Learning generalizable and identity-discriminative representations for face anti-spoofing. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–19, 2020.

[14] Andrea Lagorio, Massimo Tistarelli, Marinella Cadoni, Clinton Fookes, and Sridha Sridharan. Liveness detection based on 3d face shape analysis. In *2013 International Workshop on Biometrics and Forensics (IWBF)*, pages 1–4. IEEE, 2013.

[15] Sushil Bhattacharjee and Sébastien Marcel. What you can't see can help you-extended-range imaging for 3d-mask presentation attack detection. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7. IEEE, 2017.

[16] Lin Sun, WaiBin Huang, and MingHui Wu. Tir/vis correlation for liveness detection in face recognition. In *International Conference on Computer Analysis of Images and Patterns*, pages 114–121. Springer, 2011.

[17] Akshay Agarwal, Daksha Yadav, Naman Kohli, Richa Singh, Mayank Vatsa, and Afzel Noore. Face presentation attack with latex masks in multispectral videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 81–89, 2017.

[18] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15:42–55, 2019.

[19] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020.

[20] Sooyeon Kim, Yuseok Ban, and Sangyoun Lee. Face liveness detection using a light field camera. *Sensors*, 14(12):22471–22499, 2014.

[21] Ramachandra Raghavendra, Kiran B Raja, and Christoph Busch. Presentation attack detection for face recognition using light field camera. *IEEE Transactions on Image Processing*, 24(3):1060–1075, 2015.

[22] Zhe Ji, Hao Zhu, and Qing Wang. Lfhog: A discriminative descriptor for live face detection from light field image. In *2016 IEEE international conference on image processing (ICIP)*, pages 1474–1478. IEEE, 2016.

[23] Alireza Sepas-Moghaddam, Luis Malhadas, Paulo Lobato Correia, and Fernando Pereira. Face spoofing detection using a light field imaging framework. *IET Biometrics*, 7(1):39–48, 2017.

[24] Xiaohua Xie, Yan Gao, Wei-Shi Zheng, Jianhuang Lai, and Junyong Zhu. One-snapshot face anti-spoofing using a light field camera. In *Chinese Conference on Biometric Recognition*, pages 108–117. Springer, 2017.

[25] Valeria Chiesa and Jean-Luc Dugelay. Advanced face presentation attack detection on light field database. In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–4. IEEE, 2018.

[26] Mengyang Liu, Hong Fu, Ying Wei, Yasar Abbas Ur Rehman, Lai-man Po, and Wai Lun Lo. Light field-based face liveness detection with convolutional neural networks. *Journal of Electronic Imaging*, 28(1):013003, 2019.

[27] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection with component dependent descriptor. In *2013 International Conference on Biometrics (ICB)*, pages 1–6. IEEE, 2013.

[28] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015.

[29] Jiangwei Li, Yunhong Wang, Tieniu Tan, and Anil K Jain. Live face detection based on the analysis of fourier spectra. In *Biometric technology for human identification*, volume 5404, pages 296–303. SPIE, 2004.

[30] Santosh Tirunagari, Norman Poh, David Windridge, Aamo Iorliam, Nik Suki, and Anthony TS Ho. Detection of face spoofing using visual dynamics. *IEEE transactions on information forensics and security*, 10(4):762–777, 2015.

[31] Talha Ahmad Siddiqui, Samarth Bharadwaj, Tejas I Dhamecha, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Face anti-spoofing with multifeature videolet aggregation. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1035–1040. IEEE, 2016.

[32] Xiaobai Li, Jukka Komulainen, Guoying Zhao, Pong-Chi Yuen, and Matti Pietikäinen. Generalized face anti-spoofing by detecting pulse from face videos. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 4244–4249. IEEE, 2016.

[33] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.

[34] Keyurkumar Patel, Hu Han, and Anil K Jain. Cross-database face antispoofing with robust feature representation. In *Chinese Conference on Biometric Recognition*, pages 611–619. Springer, 2016.

[35] Klaus Kollreider, Hartwig Fronthaler, Maycel Isaac Faraj, and Josef Bigun. Real-time face detection and motion analysis with application in "liveness" assessment. *IEEE Transactions on Information Forensics and Security*, 2(3):548–558, 2007.

[36] Zhenqi Xu, Shan Li, and Weihong Deng. Learning temporal features using lstm-cnn architecture for face anti-spoofing. In *2015 3rd IAPR asian conference on pattern recognition (ACPR)*, pages 141–145. IEEE, 2015.

[37] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4680–4689, 2019.

[38] Peng Zhang, Fuhao Zou, Zhiwen Wu, Nengli Dai, Skarpness Mark, Michael Fu, Juan Zhao, and Kai Li. Feathernets: Convolutional neural networks as light as feather for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[39] Alireza Sepas-Moghaddam, Luis Malhadas, Paulo Lobato Correia, and Fernando Pereira. Face spoofing detection using a light field imaging framework. *IET Biometrics*, 7(1):39–48, 2018.

[40] Alireza Sepas-Moghaddam, Fernando Pereira, and Paulo Lobato Correia. Light field-based face presentation attack detection: reviewing, benchmarking and one step further. *IEEE Transactions on Information Forensics and Security*, 13(7):1696–1709, 2018.

[41] Alireza Sepas-Moghaddam, Mohammad A Haque, Paulo Lobato Correia, Kamal Nasrollahi, Thomas B Moeslund, and Fernando Pereira. A double-deep

spatio-angular learning framework for light field-based face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4496–4512, 2019.

[42] Alireza Sepas-Moghaddam, Ali Etemad, Paulo Lobato Correia, and Fernando Pereira. A deep framework for facial emotion recognition using light field images. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019.

[43] Alireza Sepas-Moghaddam, Ali Etemad, Fernando Pereira, and Paulo Lobato Correia. Long short-term memory with gate and state level fusion for light field-based face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1365–1379, 2020.

[44] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, Pat Hanrahan, et al. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR*, 2(11):1–11, 2005.

[45] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[47] Mindspore. `https://www.mindspore.cn/`.

[48] Information technology-presentation attack detection—part 3: Testing, reporting and classification of attacks. Standard ISO/IEC 30107- 3:2017, International Organization for Standardization, September 2017.